ISSN 0006-3509, Biophysics, 2006, Vol. 51, No. 4, pp. 523–526. © Pleiades Publishing, Inc., 2006.
Original Russian Text © R.N. Nurtdinov, A.D. Neverov, D.B. Mal'ko, I.A. Kosmosdem'yanskii, E.O. Ermakova, V.E. Ramenskii, A.A. Mironov, M.S. Gelfand, 2006, published in Biofizika, 2006, Vol. 51, No. 4, pp. 589–592.

MOLECULAR BIOPHYSICS

EDAS—Alternatively Spliced Human Gene Database

R. N. Nurtdinov[a], A. D. Neverov[b], D. B. Mal'ko[b], I. A. Kosmosdem'yanskii[c], E. O. Ermakova[a,e], V. E. Ramenskii[d], A. A. Mironov[a,b,e], and M. S. Gelfand[a,b,e]

[a]*Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, 199992 Russia*
[b]*GOSNIIGENETIKA State Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, 117545 Russia*
[c]*Biological Faculty, Lomonosov Moscow State University, Moscow, 199992 Russia*
[d]*Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia*
[e]*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia*
E-mail: gelfand@iitp.ru

Received October 2, 2005; in final form, January 23, 2006

Abstract—EDAS, an alternatively spliced human gene database, contains data on alignment of proteins, mRNAs, and ESTs. For 8324 human genes, the database contains information on all observed exons and introns and also elementary alternatives formed therefrom. The database allows one to filter the output data by varying the cutoff threshold according to the significance level. The database is available at http://www.genebee.msu.ru/edas/.

DOI: 10.1134/S0006350906040026

Key words: alternative splicing, elementary alternative, intron, exon, EST

INTRODUCTION

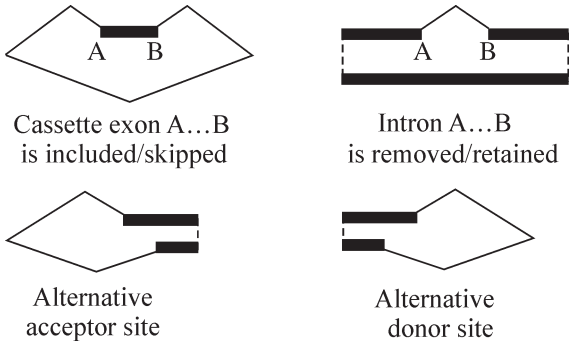
Large-scale EST sequencing showed that alternative splicing is one of the main mechanisms of maintaining the protein diversity in eukaryotes [1–3]. According to current estimates, the fraction of alternatively spliced genes in human or mouse ranges between 40 and 60% [4–6]. The insufficient purity of EST data (poor sequencing quality, mRNA isolation errors, impurities of incompletely spliced mRNAs from the cell nucleus, sequencing of single misspliced mRNAs, and mRNA cloning artifacts) casts doubt on the reliability of exon–intron structures observed in EST. However, exons and introns observed simultaneously in several ESTs obtained at different laboratories can already be considered quite a reliable object. The purpose of the creation of the EDAS database was to collect information on alignment of all known protein, mRNA, and EST sequences and combine the obtained exon–intron structures into a unit.

CREATION OF THE DATABASE

All the DNA, mRNA, EST, and protein sequences were taken from the NCBI databank [7] (MapViewer, UniGene, Entrez). The EST and mRNA sequences were aligned with genomic DNA using the program ProEST [1]; the protein sequences, using the program Pro-Frame [8].

All the observed exon–intron structures were combined. Identical internal exons and introns were grouped into a single object—internal DNA exon and DNA intron. All the first exons that coincided in donor site with any of the internal exons and were entirely within it were ignored. Similarly, all the last exons whose acceptor sites coincided with the acceptor site of any of the internal exons were ignored, too. From the rest of exons, the longest were chosen to generate the first and the last DNA exons.

For each DNA exon and DNA intron, the significance level was indicated. If a given exon (intron)



Classification of simple elementary alternatives

was observed in a protein or an mRNA, its significance level was considered maximal. For an exon (intron) observed only in an EST, the significance level was taken to be the number of different libraries in which the given exon (intron) was observed.

On the basis of sets of DNA exons and introns for each gene, a splicing graph was constructed. Nodes of this graph were splicing sites, and its edges were exons and DNA. A required significance level could be specified, and edges with the significance level below the given value could be removed from this graph.

To obtain elementary alternatives, we found all such pairs of nodes in the splicing graph that met two conditions: there is more than one path between the nodes and all these paths have no common nodes. Such a pair of nodes forms an elementary alternative. Depending on the number of paths and their lengths, several types of elementary alternatives were distinguished: cassette exon, retained intron, alternative acceptor site, and alternative donor site (figure).

Based on information on tissues and organs, developmental stages, and tumor state in EST libraries, the EDAS database enables one to isolate genes that are primarily expressed in a given set of tissues and also determine tissue-specific elementary alternatives. To determine the tissue specificity of an elementary alternative, the distribution of ESTs from tissues of interest in each of the paths forming the alternative is studied. An alternative is tissue-specific if, for a certain path, there is a statistically significant (according to the F test [9]) difference of the distribution of the ESTs studied from the distribution that is uniform over the entire alternative.

WORK WITH THE DATABASE

Chromosomal arrangement and information on alternative splicing of the genes studied are available at http://www.genebee.msu.ru/edas/data.cgi?Organism=Hs.

The distribution of the number of elementary alternatives of each type versus significance level is presented in the table.

Text search using the identifiers UniGene, GenBank, and LocusLink; short name search; and search by word in gene name are available at http://www.genebee.msu.ru/edas/search.html.

Tissue search for entire gene expression or for elementary alternatives is available at http://www.genebee.msu.ru/edas/tissues.html.

Distribution of the number of elementary alternatives of each type versus significance level

Type of alternative	Protein	mRNA	Number of EST libraries					Total
			5	4	3	2	1	
Cassette exon	1782	2236	871	375	671	1700	7036	14671
Alternative acceptor	668	882	384	175	370	905	5136	8520
Alternative donor	281	510	188	105	196	520	3272	5072
Retained intron	29	13	18	6	8	20	0	94
Total for current significance level	2760	3641	1461	661	1245	3145	15444	28357
Total	2760	6401	7862	8523	9768	12913	28357	

USE OF THE DATABASE

The database was used to study the function and evolution of alternative splicing. In particular, the algorithm IsoformCounter for generating alternative isoforms from the obtained repertoire of elementary alternatives has been developed [10]. This algorithm allows one to consider, from a single standpoint, alternative splicing of genes with substantially different levels of expression and eliminate missplicing results. This enabled one to show that alternative splicing of mRNAs of ribosomal proteins, which are expressed in a large amount and generate a great number of seeming elementary alternatives, is actually less frequent than on the average over the genome. Another functional category is which alternative splicing is relatively avoided is signal transmission related to small GTPases. On the other hand, a relative increase was revealed in the frequency of alternative splicing of genes in the functional category of DNA replication and chromosomal cycle as well as among genes whose products are involved in protein–protein interactions.

Comparison of alternatively spliced genes of human and mouse showed that alternative exons and splicing sites are much less conserved than constitutive ones [11, 12]. Comparison of nucleotide substitutions in alternative and constitutive regions demonstrated that the nonsynonymous substitution rate in alternative regions is relatively higher, which may suggest the lack of stabilizing selection or the existence of positive selection in these regions [13]. A similar observation was made in analyzing single-nucleotide polymorphisms in the human genome [14].

These observations together with similar works of other researchers [15–20] allowed us to formulate the concept of alternative splicing as a "playground of evolution" [21]. Young alternative regions enable testing of possible new protein functions without losing the old ones. If an additional protein segment is functionally useful, the fraction of the isoform involving this segment can be increased by small adjustments in regulation of alternative splicing. Within the framework of this model, observations of an increase in the molecular evolution rate in alternative regions are well fit.

ACKNOWLEDGMENTS

We thank A. Kazakov, A. Fedorov, and I. Erokhin for assistance in processing of EST tissue classification data; V. Makeev for valuable discussions and help in the work at the text; and also P. Novichkov for advice on programming.

This work was supported by the Howard Hughes Medical Institute, the Russian Foundation for Basic Research (project no. 04-04-49440), the Program of the Russian Academy of Sciences "Molecular and Cell Biology", and the Russian Science Support Foundation (M.S.G.).

REFERENCES

1. A. A. Mironov, J. W. Fickett, and M. S. Gelfand, Genome Res. **9**, 1288 (1999).
2. D. Brett, J. Hanke, G. Lehmann, *et al.*, FEBS Lett. **474**, 83 (2000).
3. E. V. Kriventseva, I. Koch, R. Apweiler, *et al.*, Trends Genet. **19**, 124 (2003).
4. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature **409**, 860 (2001).
5. D. Brett, H. Pospisil, J. Valcarcel, *et al.*, Nature Genet. **30**, 29 (2002).
6. FANTOM2 Consortium and the RIKEN GSC Genome Exploration Group Phase I & II Team, Analysis of the mouse transcriptome based upon functional annotation of 60,770 full length cDNAs, Nature **420**, 563 (2002).
7. D. L. Wheeler, D. M. Church, R. Edgar, *et al.*, Nucl. Acids Res. **32**, D35 (2004).
8. A. A. Mironov, P. S. Novichkov, and M. S. Gelfand, Bioinformatics **17**, 13 (2001).
9. A. Agresti, Stat. Sci. **7**, 131 (1992).
10. A. D. Neverov, I. I. Artamonova, R. N. Nurtdinov, *et al.*, BMC Bioinformatics **6**, 266 (2005).
11. R. N. Nurtdinov, A. A. Mironov, and M. S. Gel'fand, Biofizika **47**, 197 (2002).
12. R. N. Nurtdinov, I. I. Artamonova, A. A. Mironov, and M.S. Gelfand, Hum. Mol. Genet. **12**, 1313 (2003).

13. E. O. Ermakova, *Proceedings of the Second Moscow Conference on Computational Molecular Biology «MCCMB'2005»* (Moscow, 2005).

14. V. E. Ramensky, A. D. Neverov, R. N. Nurtdinov, *et al.*, *Proceedings of the Second Moscow Conference on Computational Molecular Biology «MCCMB'2005»* (Moscow, 2005).

15. T. A. Thanaraj, F. Clark, and J. Muilu, Nucl. Acids Res. **31**, 2544 (2003).

16. B. Modrek and C. J. Lee, Nature Genet. **34**, 177 (2003).

17. G. W. Yeo, E. Van Nostrand, D. Holste, *et al.*, Proc. Acad. Sci. USA **102**, 2850 (2005).

18. Y. Xing and C. Lee, Proc. Acad. Sci. USA **102**, 13526 (2005).

19. Y. Xing and C. Lee, Bioinformatics **21**, 3701 (2005).

20. B. P. Cusack and K. H. Wolfe, Mol. Biol. Evol. **22**, 2198 (2005).

21. M. S. Gel'fand, Russ. Zh. VICh/SPID Rodstv. Probl. **8**, 16 (2004).